University of Minnesota, Duluth

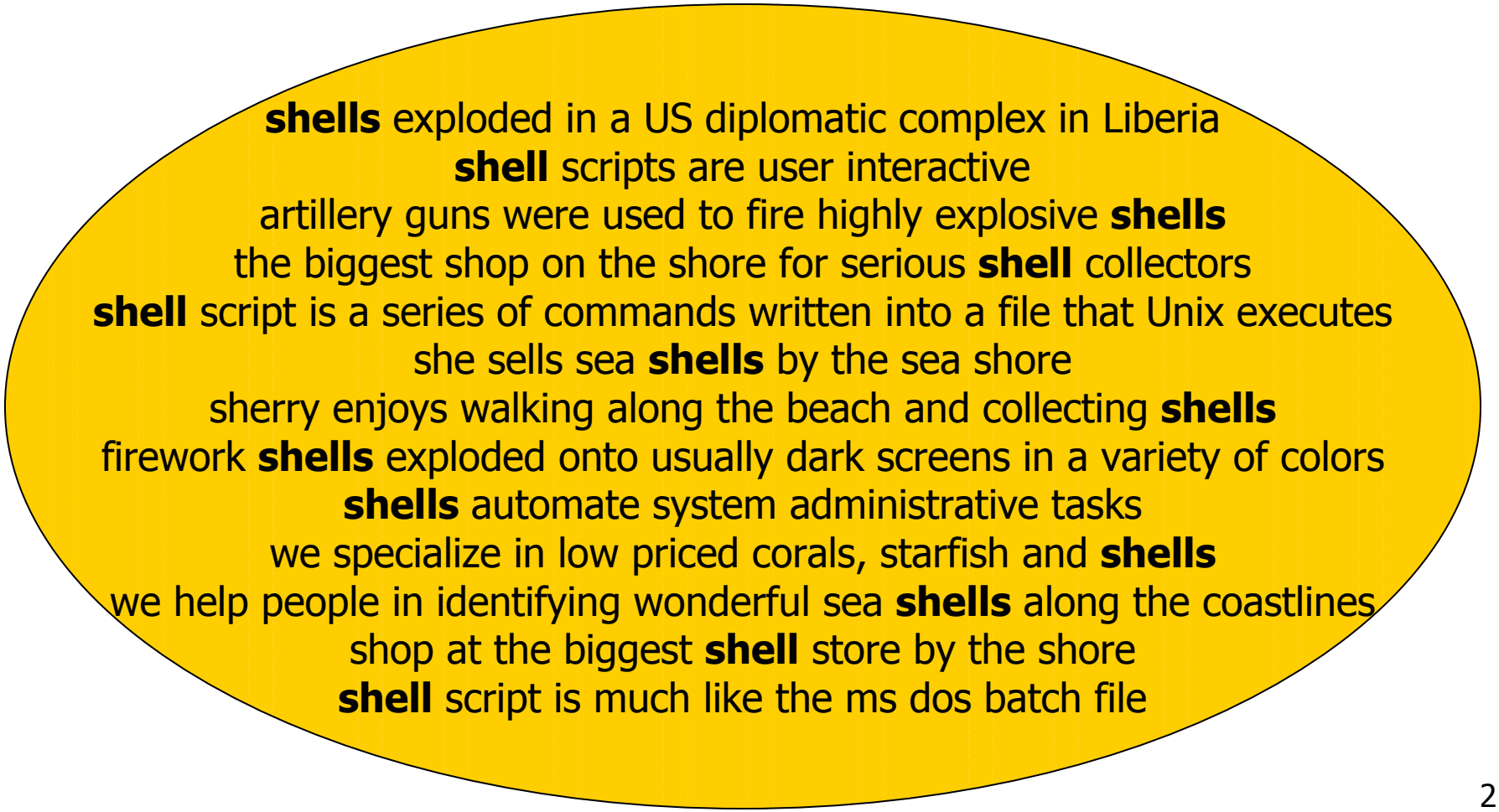
Discriminating Among Word Meanings by Identifying Similar Contexts

http://senseclusters.sourceforge.net

Amruta Purandare
Advisor: Dr. Ted Pedersen
July 27, 2004

Research Supported by National Science Foundation
Faculty Early Career Development Award (#0092784)

Word Sense Discrimination


shells exploded in a US diplomatic complex in Liberia
shell scripts are user interactive
artillery guns were used to fire highly explosive **shells**
the biggest shop on the shore for serious **shell** collectors
shell script is a series of commands written into a file that Unix executes
she sells sea **shells** by the sea shore
sherry enjoys walking along the beach and collecting **shells**
firework **shells** exploded onto usually dark screens in a variety of colors
shells automate system administrative tasks
we specialize in low priced corals, starfish and **shells**
we help people in identifying wonderful sea **shells** along the coastlines
shop at the biggest **shell** store by the shore
shell script is much like the ms dos batch file

shells exploded in a US diplomatic complex in Liberia
firework **shells** exploded onto usually dark screens in a variety of colors
artillery guns were used to fire highly explosive **shells**

sherry enjoys walking along the beach and collecting **shells**
we specialize in low priced corals, starfish and **shells**
we help people in identifying wonderful sea **shells** along the coastlines
shop at the biggest **shell** store by the shore
she sells sea **shells** by the sea shore
the biggest shop on the shore for serious **shell** collectors

shell script is much like the ms dos batch file
shell script is a series of commands written into a file that Unix executes
shell scripts are user interactive
shells automate system administrative tasks

Our Approach

- Strong Contextual Hypothesis

Sea Shells => (sea, beach, ocean, water, corals)

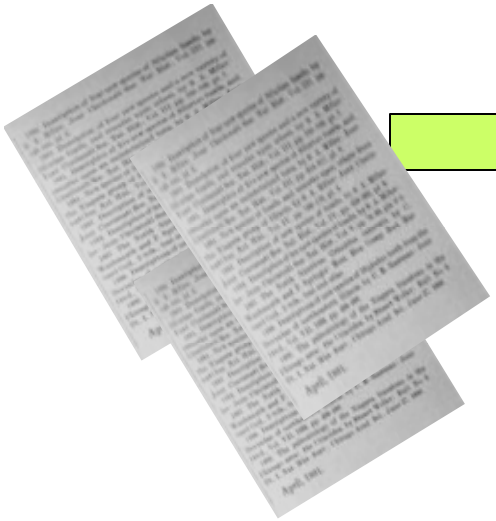
Bomb Shells => (kill, attack, fire, guns, explode)

Unix Shells => (machine, OS, computer, system)

- Corpus—Based Unsupervised Clustering

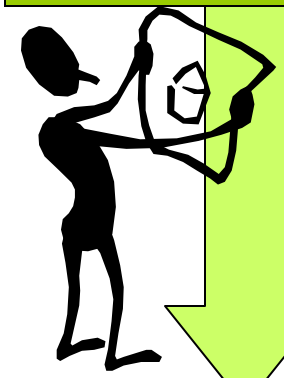
- Knowledge—Lean
  - Portable – Other languages, domains
  - Scalable – Large Raw Text
  - Adaptable – Fluid Word Meanings

Methodology



Feature Selection

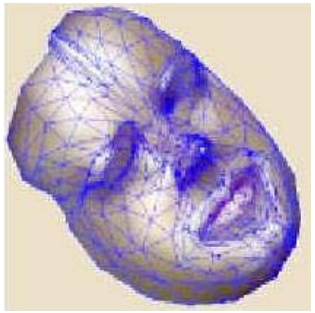
Context Representation



Measuring Similarities

Clustering

Evaluation



Surface Lexical Features

- Unigrams

  in today's world the scallop is a popular design in architecture and is well known as the shell gasoline logo

- Bigrams

  she sells sea shells on the sea shore

- Co-occurrences

  bivalve shells are mollusks with two valves joined by a hinge

  shells can decorate an aquarium

Local Training

Pectens or Scallops are one of the few bivalve **shells** that actually swim. This is accomplished by rapidly opening & closing their valves, sending the shell backward.

Fire marshals hauled out something that looked like a rifle with tubes attached to it, along with several bags of bullets and **shells**.

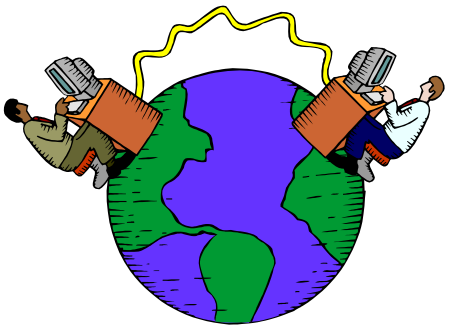
If you hear a snapping sound when you're in the water, chances are it is the sound of the valves hitting together as it opens and shuts its **shell**.

Teenagers tried to make a bomb or some kind of homemade fireworks by taking the bullets and shotgun **shells** apart and collecting the black powder.

Bivalve **shells** are mollusks with two valves joined by a hinge. Most of the 20,000 species are marine including clams, mussels, oysters and scallops.

There was an explosion in one of the **shells**, it flamed over the top of the other shells and sealed in the fireworks, so when they ignited, it made it react like a pipe bomb."

These edible oysters are the most commonly known throughout the world as a popular source of seafood. The **shell** is porcelaneous and the pearls produced from these edible oysters have little value.



Global Training

U.S. researchers said sea shells may be the product of a geological accident that flooded ancient oceans with calcium, thereby diversifying marine life. Researchers at the U.S. Geological Survey have found the amount of calcium in sea water shot up between the end of the Proterozoic era (about 544 million years ago) and the early Cambrian period (515 million years ago). This increase, they suggested, allowed soft-bodied marine organisms to create hard shells or body parts from the calcium minerals. The researchers studied the chemical composition of liquids trapped in the cavities of salty rocks called halites, which provide samples of prehistoric oceans.

John Kerry is a man who knows how to keep a secret. The Democratic White House hopeful was so obsessed with making sure the name of John Edwards, his vice presidential running mate, remained under wraps until the announcement that he had vendors who printed up placards and T-shirts sign a non-disclosure agreement. Kerry himself telephoned his plane charter company at 6 p.m. on Monday night to let them in on his decision in time to have the red, white and blue aircraft's decal changed to read "Kerry-Edwards A Stronger America." Edwards did not travel to Pittsburgh to attend the rally at which his name was announced, which also might have alerted the media. After months of speculation, first reports began emerging less than 90 minutes before Kerry made his public announcement at 9 a.m.

1st Order Context Vectors

C1: if she sells shells by the sea shore, then the shells she sells must be sea shore shells and not firework shells

C2: store the system commands in a unix shell and invoke csh to execute these commands

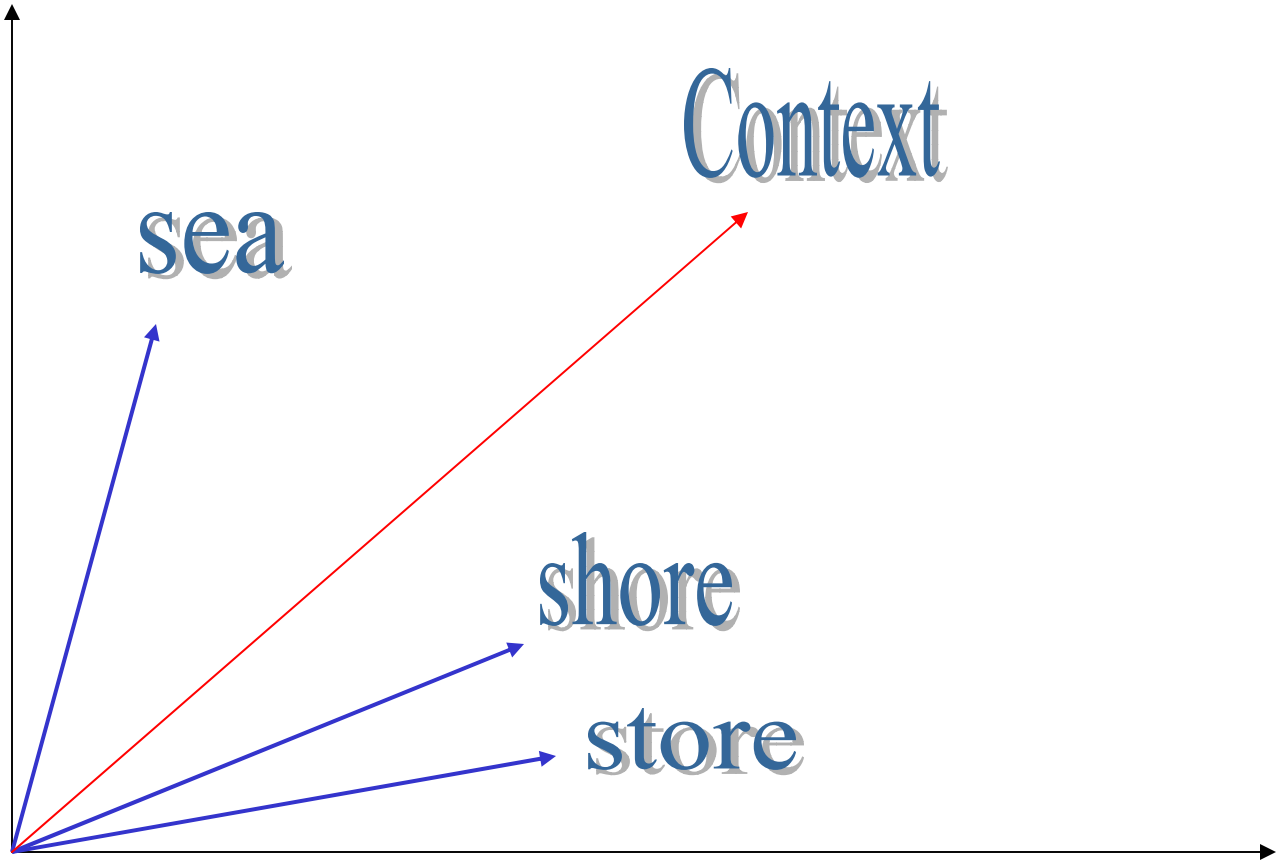
	sea	shore	system	execute	firework	unix	commands
C1	2	2	0	0	1	0	0
C2	0	0	1	1	0	1	2

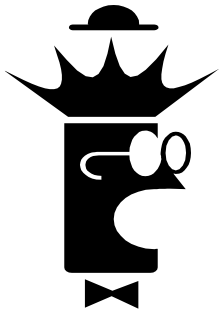
2nd Order Context Vectors

The largest shell store by the sea shore

	Sells	Water	North-West	Sandy	Bombs	Sales	Artillery
Sea	18.5533	3324.98	30.520	51.7812	8.7399	0	0
Shore	0	0	29.576	136.0441	0	0	0
Store	134.5102	205.5469	0	0	0	18818.55	0
O2 context	51.021	1176.84	20.032	62.6084	2.9133	6272.85	0

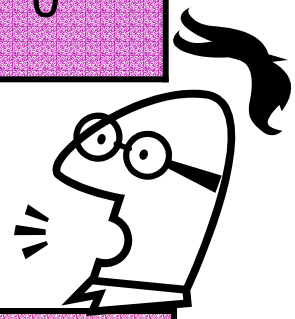
2nd Order Context Vectors





Similarity 0 ???

Kill	Murder	Destroy	Fire	Shoot	Missile	Weapon
2.53	0	1.28	0	3.24	0	28.72
0	4.21	0	0.92	0	52.27	0



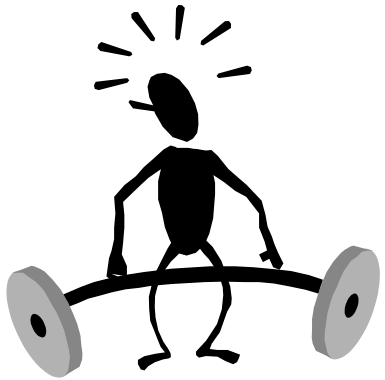
High Similarity ???

Burn	CD	Fire	Pipe	Bomb	Command	Execute
2.56	1.28	0	72.7	0	2.36	19.23
34.2	0	22.1	46.2	14.6	0	17.77

Latent Semantic Analysis

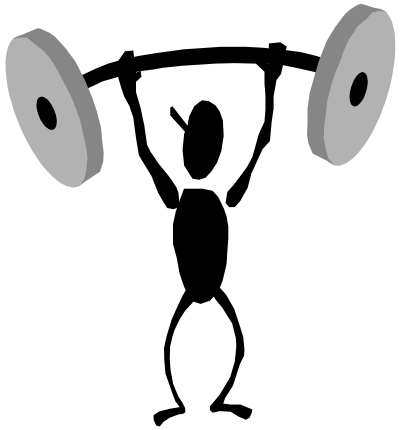
- Singular Value Decomposition
- Resolves Polysemy and Synonymy
- Captures Conceptual Similarities
- Converts Word Space to Semantic Space

Before SVD …



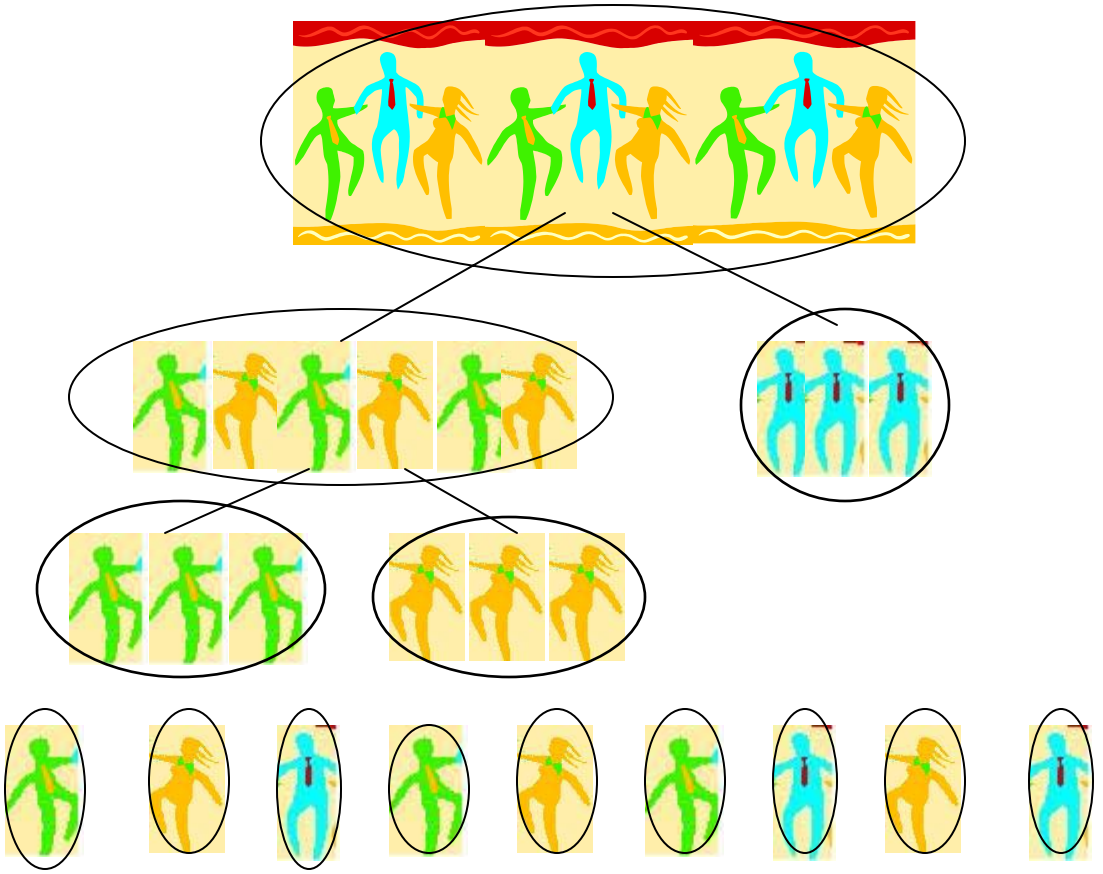
Computer	Kill	Unix	Pipe	Bomb	Guns	Murder	Execute	Fire	Shoot	Machine
2	0	1	0	0	0	0	2	0	0	3
0	2	0	1	2	3	1	3	2	0	2
0	3	0	2	2	3	2	0	0	2	3
0	1	0	0	2	1	0	1	2	0	0
0	0	0	0	1	2	2	2	2	0	0
2	0	2	2	0	0	0	3	0	0	2
2	1	3	3	0	0	0	2	0	0	3

After SVD …



Computer	Kill	Unix	Pipe	Bomb	Guns	Murder	Execute	Fire	Shoot	Machine
1.48	0.36	1.59	1.45	-0.03	0.021	-0.012	1.713	-0.09	0.028	2.073
0.222	2.12	0.24	1.34	2.11	2.899	1.587	2.022	1.58	0.75	2.181
0.102	2.11	0.11	1.24	2.16	2.93	1.608	1.906	1.61	0.761	2.04
-0.306	1	-0.32	0.289	1.11	1.512	0.834	0.5712	0.85	0.386	0.552
-0.27	1.25	-0.28	0.45	1.35	1.84	1.013	0.812	1.03	0.471	0.813
1.934	0.42	2.07	1.87	-0.08	-0.03	-0.05	2.2	-0.15	0.02	2.67
2.292	0.59	2.45	2.262	-0.01	0.075	0.005	2.675	-0.11	0.054	3.23

Clustering



- UPGMA
  - Hierarchical : Agglomerative
- Repeated Bisections
  - Hybrid : Divisive + Partitional

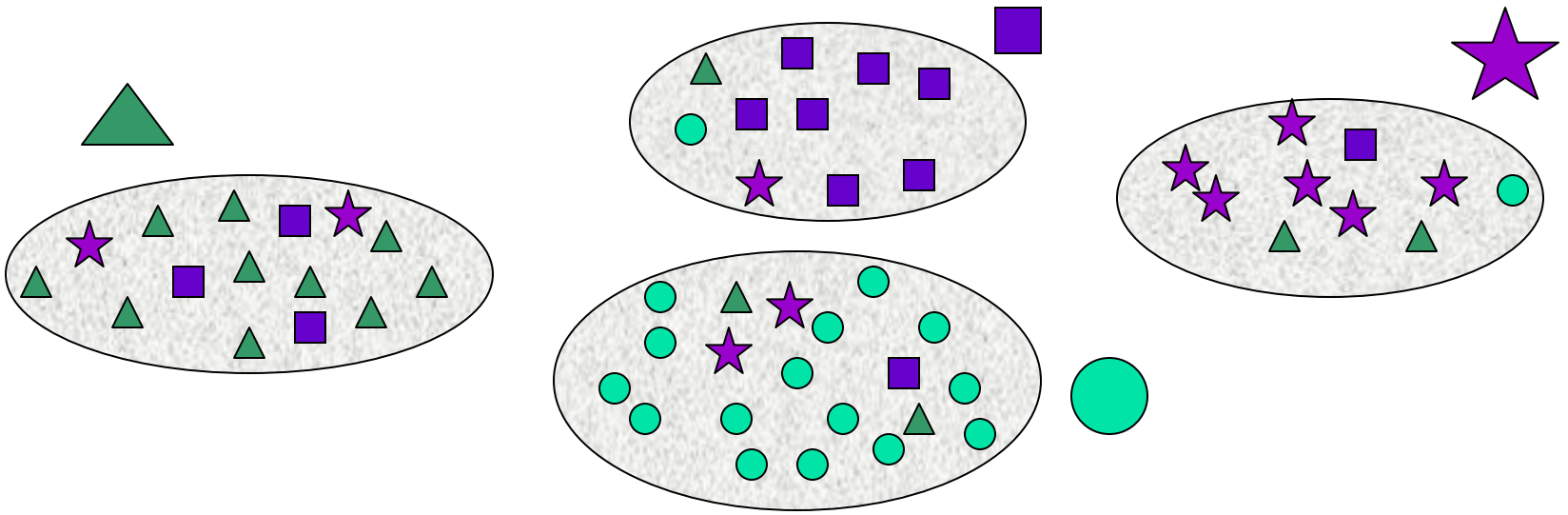
Similarity Coefficients

Matching	$(X \sqcap Y)$
Dice	$\frac{2\,(X \sqcap Y)}{\lvert X\rvert + \lvert Y\rvert}$
Jaccard	$\frac{(X \sqcap Y)}{\lvert X \cup Y\rvert}$
Overlap	$\frac{(X \sqcap Y)}{\min(\lvert X\rvert,\lvert Y\rvert)}$
Cosine	$\frac{(X \sqcap Y)}{\sqrt{\lvert X\rvert * \lvert Y\rvert}}$

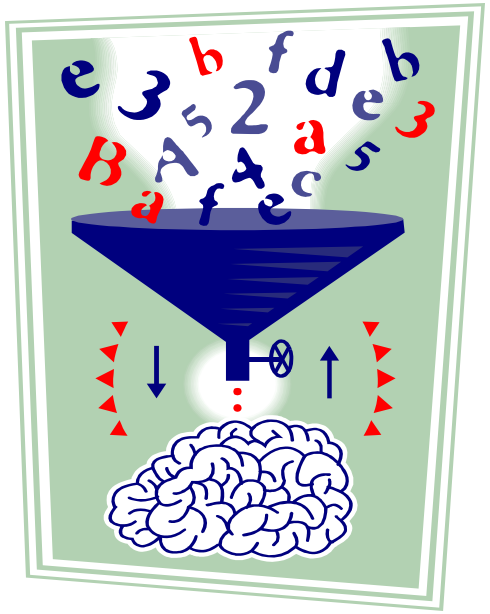
Evaluation

	▲	■	★	●	
C1	**10**	3	2	0	15
C2	1	**7**	1	1	10
C3	2	1	**6**	1	10
C4	2	1	2	**15**	20
	15	12	11	**17**	55

Accuracy=38/55=0.69



Data



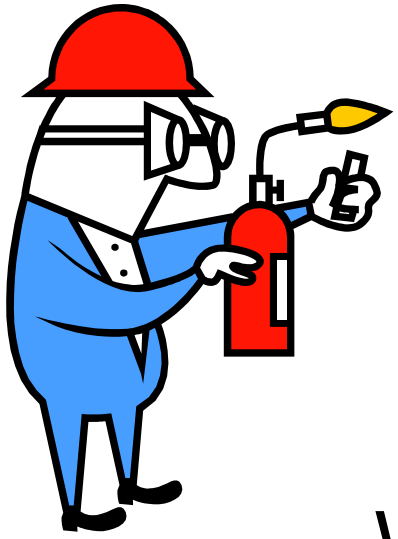
- Line, Hard, Serve
  - 4000+ Instances / Word
  - 60:40 training:test Split
  - 3-5 Senses / Word
- SENSEVAL-2
  - 72 words = 28 V + 29 N + 15 A
  - Approx. 50-100 Test, 100-200 Training
  - 8-12 Senses / Word
- Associated Press Worldstream Newswire
  - Nov 1994-June 2002 by LDC, U Penn
  - 539,665,000 words

Experiments

- Training
  - Local *versus* Global
- Features
  - Unigrams, Bigrams, Co-occurrences
- Context Representations
  - 1st order *versus* 2nd order
- Clustering
  - RBR *versus* UPGMA

Conclusions

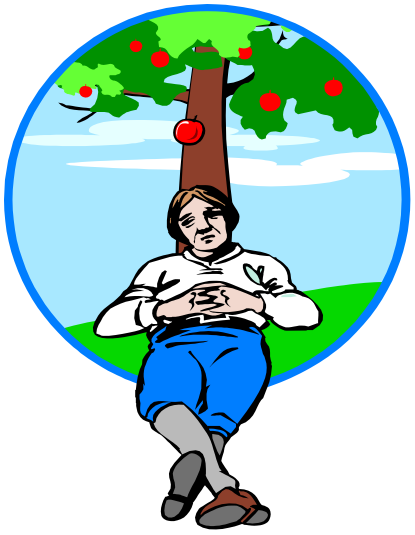
- Smaller Data
  - 2nd Order + RBR
- Larger Local Data
  - 1st Order + UPGMA
- Global Training
  - Helps 1st Order Bigrams, UPGMA
  - Overall Local Training Better
- Corpus + Dictionary
  - Helps only with smaller training data



Applications

- Word Sense Disambiguation
- Synonymy Identification
- Email/News Classification
- Name Discrimination
- Ontology Acquisition
- Text Summarization

Contributions



- Systematic Comparison
  - Pedersen & Bruce (1997)
  - Schütze (1998)
- Discrimination Parameters
  - Features
  - Context Representations
  - Clustering Approaches

Contributions

- Training Variations
  - Local
  - Global
- Relative Comparison
  - Raw Corpus
  - Corpus + Dictionary
- Software
  - *http://senseclusters.sourceforge.net*

Thank You !

Visit our Demo at Table #212